Deduplication Policy

Alaska Immunization Program
3601 C Street, Suite 540
Anchorage Alaska 99503
Tel 907-269-0312
Toll Free 866-702-8725
Fax 907-562-7802
vactrak@alaska.gov

IRMS (Information Registry Management System) denotes the health organization composed of all the facilities, vaccinators and physicians covered by your VacTrAK application.

Deduplication is one of the most important tasks of VacTrAK. As records come in from electronic data systems from around the state, VacTrAK is responsible for deduplicating not just immunizations for a given patient, but deduplicating patient records as well.

Automatic deduplication

VacTrAK has an efficient and detailed automatic deduplication process which will merge two distinct records when VacTrAK determines the two records are the same patient. This merge process is based on a weighted comparison of the demographic data contained in the two records, including but not limited to:

- Patient name
- Date of birth
- Social Security number
- Medicaid number
- Guardian information
- Mother's maiden name
- Address

For these reasons, it is important that providers engaging in data exchange with VacTrAK include as many of these fields as possible to aid in the deduplication process.

The automatic deduplication process runs daily at 6:00 AM. It generally takes less than an hour and will run in the background even while users are accessing VacTrAK via the internet. The automatic deduplication description published by STC, the software vendor for VacTrAK, is included (Appendix A).

Manual Deduplication

If VacTrAK's automatic process is unable determine whether or not two records are the same person or different persons, the records will be staged in Manual Deduplication to

be examined by VacTrAK Support. This most often occurs with name or date of birth typos, or where there is too little information in the demographic record for VacTrAK to determine whether the patients are the same or separate. These will be reviewed on a daily basis by VacTrAK Support.

VacTrAK Support will merge two records where the first name, last name and date of birth all match (Appendix B). If a name or date of birth is slightly different, but other components like Social Security number, Medicaid number, and/or address match a determination will be made by VacTrAK Support as to whether the two records shall be merged or marked as separate.

Extra precautions are taken for records of patients through 18 years of age because the inappropriate merging of two pediatric records is more problematic than for two adult records.

Once a decision is made in manual deduplication any further imported data for the patient from the originating IRMS will abide by this decision. The record will not re-appear in manual deduplication for review each time the record is imported.

IRMS Identified Internal Patient Deduplication

Each IRMS is responsible for reporting to VacTrAK when records in their system are identified as duplicates or bad merges.

VacTrAK Support will accept deduplication reports from each IRMS for their patients and modify VacTrAK appropriately. VacTrAK is not responsible for merging patients within the IRMS's dataset.

VacTrAK User Identified Patient Deduplication

VacTrAK users may identify possible duplicates in VacTrAK and may report the duplicates to VacTrAK Support for resolution by using the *Report Possible Duplicates* button at the bottom right of the patient search page. The user must submit a clear and comprehensive comment explaining why the two are suspected duplicates. The user reporting the duplicate should include their clinic's name and telephone number in the comments section.

VacTrAK Support will monitor the user-reported duplicates on a daily basis and may contact providers in order to make a determination. Once a decision has been made, VacTrAK Support will inform the submitting user of the outcome.

User-Reported Bad Merges

In the automated deduplication process, VacTrAK may erroneously merge two distinct patient records which should be kept separate. If a provider identifies possible bad merges, they must contact VacTrAK Support (269-0312 or 1-866-702-8725). The user reporting the bad merge must provide clear and comprehensive reasoning explaining why the single record is a suspected bad merge.

Ambiguous ID

In the event that VacTrAK receives two distinct records for two distinct patients from a single provider with the same medical record number, the records will automatically be flagged for manual review. The IRMS is responsible for preventing non-unique IDs for patient records within their system. The IRMS will be responsible for reviewing the records flagged for ambiguous ID by VacTrAK Support.

Appendix A.

Deduplication Records, Rules Defined, and More

An excerpt from Scientific Technologies Corporation. (2008). *Immunization Management System Administration Guide* (pp 6-1 to 6-23). Tucson: STC.

DEDUPLICATION RECORDS, RULES DEFINED, AND MORE

The registry is designed to minimize the number of duplicate patient and vaccination records within it; however, it is still possible for duplicates to occur when using both the Online Interface or Batch Load methods.

Although the Online Interface forces the user to perform a thorough search for a patient, duplications are still possible due to the following while performing a search:

- Variations in the name spelling
- Address changes
- Typographical errors

Overall, any difference between how the patient was originally entered and how it's being searched for, can result in the system not locating a match and thus, a duplicate patient is entered.

The "Search" process requires a minimum amount of search criteria to be entered. This information is fed into a sophisticated **search algorithm** that tries many different combinations of the information against the necessary database tables. If the patient is still not found, the user can enter the patient as "new."

Batch Loads to the Registry are sent by organizations that independently maintain their own database; hence, their own listing of patient records. Since they are maintaining their own database of records, it is very possible to send duplicates. These duplications must be handled centrally at the registry as they come into the system.

There are two reasons a duplicate can be sent by batch:

- If there is a duplicate in the reporting database, the duplicates will be sent to the registry.

- When a patient goes to another clinic that is not served by the same database, a duplicate record is created and will be sent to the central registry.

When the records are sent via Batch Load, the data is stored in a "holding" table called the "Pre-Reserve" table. The data stays in this table until the "Automatic Duplicate Identification Procedure," also referred to as the "Deduplication" process runs. The Deduplication process moves the record, one record at a time, based on the process "result." The record's destination is the "Reserve" and "Master" tables; however, depending on the Duplicate Identification "result," determines the record's destination.

DEDUPLICATION PROCESSES

Obviously, the goal is to avoid creating duplicate records for the same patient; this process is referred to as "**Deduplication**" or "**Duplicate Identification**."

There are four processes that can determine whether a record in the database is a duplicate or not. The processes are:

- **Automatic Duplicate Identification Procedure**
  - **Deterministic (Default)**
  - **Probabilistic (State configurable Property)**
- **Manual Duplicate Identification Procedure** (**Manual Deduplication**)
- **Automatic Master Duplication Procedure** (**Scanning**)
- **Manual Master Duplication Procedure**

The "Automatic Duplicate Identification Procedure" uses the "Rule-based algorithm."

The "Manual Duplicate Identification Procedure" uses the "Manual Deduplication" option. This process involves viewing both the incoming record with the database record, and making a decision; thus, a manual process.

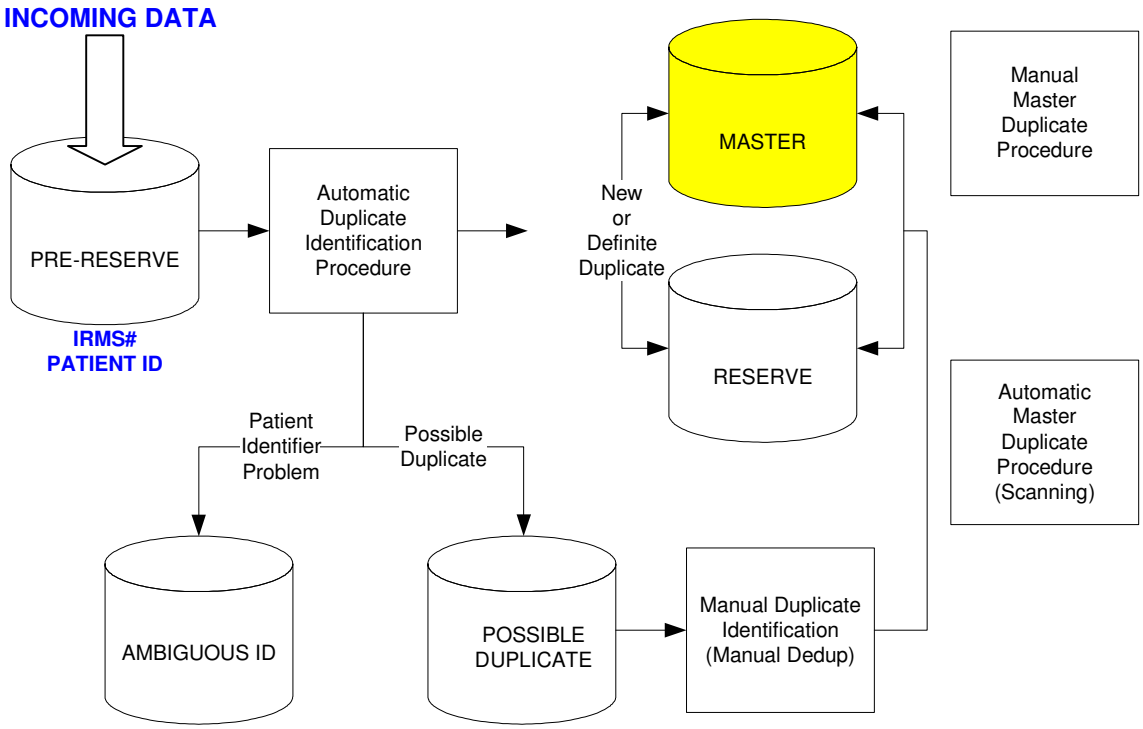
AUTOMATIC DUPLICATE IDENTIFICATION PROCESS

The Automatic Duplicate Identification Process offers two different types (Deterministic or Probabilistic); however, the identification process is the same.

The Automatic Duplicate Identification Procedure looks at the "Incoming Data" (IRMS# + PATIENT ID = PATIENT IDENTIFIER) record in the "Pre-Reserve" table and attempts to locate a match or similar record.

The illustration reveals the process, when the process may run, and the database tables a record possibly encounters.

Figure 0-1: Automatic Duplicate Identification Process



DATABASE TABLES FOR RECORD STORAGE

The database tables the record may go into are listed in the table along with an explanation as to why the record may be stored in a particular table:

Table 0-1: Database Tables for Record Storage

TABLE	DESCRIPTION
Pre-Reserve	Holds patient records waiting to be processed into the "Reserve" and "Master" tables. When the data is loaded into the "Pre-Reserve" table some of the fields are pre-processed. Both the "before and after" process record will be stored in the table. **ADDRESSES** are pre-processed to ensure conformity to postal standards and spaces are removed. **NAMES** are pre-processed in three ways: Unnecessary spaces and special characters (such as dashes in hyphenated names) are removed **NAMES** are checked to see if FIRST and MIDDLE have been swapped or if MIDDLE and LAST have been swapped. **NAMES** are checked to ensure FIRST and MIDDLE names were both entered into the FIRST name field.
Reserve	Stores the patient record exactly as it was sent from the Facility/IRMS. The Facility/IRMS database's IRMS # and Patient ID are stored in this table so that the record can be "mapped" back to the Facility/IRMS database.
Master	Stores the "latest" patient record sent from a Facility/IRMS database. Each "Reserve" record that has been mapped to this "Master" record can reference it. The "mapping" of patient "Reserve" records to patient "Master" records is used to create a composite vaccination record by combining all the vaccinations sent from Facility/IRMS databases.
Possible Duplicate	Holds incoming patient records that have possible duplicate matches in the "Reserve" table. Records only go into this table if the "Automatic Duplicate Identification" procedure could not definitely identify a duplicate record. Records in this table require "human review" using the "Manual Duplicate Identification" or Manual Deduplication.
Ambiguous ID	Holds incoming patient records that have Patient Identifier problems. For example, an IRMS has sent patient 5 twice—referred to as patient 5a and patient 5b. Patient 5a has a different name than patient 5b. Records in this

TABLE	DESCRIPTION
	table require "human review" using the "Ambiguous ID" application.

Records are moved from the "Pre-Reserve" to the "Master" table one record at a time. The journey of a record to the "Master" table depends on whether another patient record is found to be similar or to be a definite match.

The record match result will be one of the following:

- **New Record**—a match was NOT located.
- **Duplicate Record**—a match was located. For example. If the **Street Address/State/Zip Code** or **Street Address/City** matches the record in the "Reserve" table, then the end result is a Duplicate.
- **Possible Duplicate**—unsure if a match was located.
- **Patient Identifier Problem**—unique Patient Identifier is no longer unique). A separate module named "**Ambiguous ID**" is used by the Registry Administrator who will review them and decide whether the record represents the same person or not. For version 4.2, if two records come in from the same IRMS, the newest record should ALWAYS overwrite the existing record except in cases where the newest record has a null value (this field would keep what is in the original record rather than keeping the blank field).
    - The existing property “Enable Separate During Ambiguous Id” was disabled.
    - The logic was modified to process ambiguous id decisions so if the administrator has chosen to process the incoming records, the newest record will not overwrite any existing values with null (the newest record currently overwrites all existing values).
    - This new logic will execute both when the user reviews the ambiguous id records manually and selects to accept the new record and when the user

selects "Process IRMS" to process all records without looking at them individually.

- If the incoming record is intentionally nulling a column, such as an instance when the SSN was bad and the correct one is not known, this logic will prevent that from being corrected and the user will have to manually update it in IWeb.

DEDUPLICATION RULES/LOGIC

The Deduplication Rules/Logic differs between **Deterministic** and **Probabilistic**.

DETERMINISTIC (DEFAULT) RULES/LOGIC

The SIIS Deduplication logic uses rule-based algorithms to determine matches in the following order:

- Likely Matches
- Matches
- Possible Matches
- Separate Rules/Special Cases

Each type is discussed in its own separate section.

To assist in the process of determining a match, questions and answers are listed:

1. Is it possible that two different people have the same First Name and Birth Date?
    - Yes. It is possible to have two different people with the same names and birth dates, so these records would NOT be merged; however, if additional fields match, such as phone number, then it may be safe to merge the records.
2. Is it possible to have the same social security number but all the other fields do not match?
    - No. More than likely, there is a typographical error on the social security number.

3. Is it possible to have the same first name, middle name, last name, birthday, and address but have different social security numbers?

   - Yes, this is probably the same person, but the social security number was probably entered incorrectly.

4. Other factors to consider are:

   - Last names can change due to marriage, divorce, custody changes, etc.
   - Addresses change when patients move.
   - Twins will have all fields in common except for first names and possibly middle names.
   - Sons are named the same as their fathers where the record difference will only be seen in the birth date.
   - Orphanages can have several children located at the same address, so addresses and phone numbers are not significant for the record comparison.
   - Look for swapped entries; such as the First Name being placed in the Last Name field, or Middle Name in the First Name field, etc.
   - Be on the "lookout" for "systematic bad data." For example, all patients have the same social security number. The number of records coming in for "Manual Deduplication" is enormous. If this happens, the sending site must be contacted and instructed to "fix" the data before sending additional data.

RULES FOR LIKELY MATCHES

The first step of deduplication is to get a subset of "likely" matches. These are the existing patients in the database that match on any of the following field values:

- Birth Date and First Initial
- Birth Date and Last Initial
- Street Address, State, and Zip Code

- Street Address and City
- 7-digit Phone Number (excludes area code)
- Guardian SSN
- SSN

The next step is to compare each patient in that subset to the "new patient."

- If a match is found at any point, the program immediately exits and goes to the next patient.
- If a match is NOT found, the records will be flagged as a "possible match" or as "separate patients."

The following values are pre-processed:

- Addresses are pre-processed to conform to postal standards and remove spaces.
- Names are pre-processed to remove spaces and special characters (such as dashes in hyphenated names).
- Names are checked to see if First and Middle have been swapped or if Middle and Last have been swapped.
- Names are checked to see if the First and Middle names were both entered into the First Name field.

The deduplication logic examines the following fields:

- Medicaid Number
- Birth File Number
- SSN
- First Name
- Middle Name
- Last Name
- Birth Date
- Street Address, State, and Zip Code

- Street Address, City
- 7-digit phone number (excludes area code)
- Guardian SSN
- Guardian First Name
- Guardian Last Name
- Mother Maiden Name
- Gender
- Second Guardian First Name
- Second Guardian Last Name
- Second Guardian SSN
- Birth Order
- Multiple Birth Count

Additional information is determined from these fields:

Table 0-2: Additional Deduplication Logic Field Information

If Field	Quantifier	Field/Value	Then Field	Result
First Name	**EQUALS**	Baby BabyBoy BabyGirl Newborn Boy Girl	First Name	Is Identified as Special Baby Name Match
Street Address State Zip Code	**EQUALS**	Street Address City	Address	Is Identified as Address Match

If Field	Quantifier	Field/Value	Then Field	Result
Guardian Last Name or Mother's Maiden Name	**EQUALS**	Patient's Last Name	Last Name	Additional weight is given to "unique" Last Names
Patient 1's Last Name	**EQUALS**	Patient 2's Last Name, or Patient 2's Guardian's Last Name, or Patient 2's Mother's Maiden Name	Last Name	Unique Last Names is incremented by 1
Patient 1's Guardian's Last Name	**EQUALS**	Patient 2's Last Name Patient 2's Guardian Last Name Patient 2's Mother's Maiden Name	Last Name	Unique Last Names is incremented by 1
Patient 1's Mother's Maiden Name	**EQUALS**	Patient 2's Last Name Patient 2's Mother's Maiden Name Patient 2's Guardian Last Name	Last Names	Unique Last Names is incremented by 1
Patient 1's Last Name	**EQUALS**	Patient 1's Guardian Last Name Patient 1's Mother's Maiden Name	Last Names	Unique Last Name is incremented by 1
Patient 1's Guardian Last Name	**EQUALS**	Patient 1's Mother's Maiden Name	Last Names	Unique Last Name is decremented by 1
Since the unique Last Name is calculated in this manner, its value can be anywhere from 0 to 3.				

Approximate Birth Date signifies if the Birth Date is close. It is calculated as follows:

1. If the Birth Dates are less than 9 months apart, Birth Date is approximate if any of the following are true:
   - If the day and month are swapped (i.e., 10/11/2003 and 11/10/2003)
   - If the year and day are the same but the month ids different
   - If the year and month are the same but the day is different
2. If the Birth Dates are less than 10 years apart, Birth Date is approximate if the following is true:
   - If day and month match
3. First Guardian Name and Second Guardian Name are examined to check for swaps (i.e., the father was the first guardian on one records and the second guardian on the other record).
4. If both Birth Order and Multiple Birth Count are present and different and the record would have otherwise matched, send the record to Manual Deduplication.
5. If Mother Maiden Name is "ADULT" or "SELF," these values are ignored.

RULES FOR MATCHES

The **SIIS Deduplication** logic uses a "rule-based" algorithm to identify a match. These "**merge (match)**" rules are identified by a rule number. The rules, descriptions, and fields are listed in the table.

Table 0-3: Merge (Match) Rules

RULE #		FIELDS	
		Column 1	Column 2
100	Medicaid Number	N/A	N/A

RULE #		FIELDS	
		Column 1	Column 2
	and Birth File Number		
110	SSN and Birth File Number	N/A	N/A
120	First Name and at least one field from column 1 and 2	SSN Birth File Number Medicaid Number Guardian First Name	Birth Date Middle Name Unique Last Name Address Phone
121	Nickname and Gender and at least one field from column 1 and 2	SSN Birth File Number Medicaid Number	Birth Date Middle Name Unique Last Name Guardian First Name Address Phone
122	First Name and approximate Birth Date and at least one field from column 1.	SSN Birth File Number Medicaid Number	
130	Special Baby Name, Birth Date and at least two of the following fields from column 1.	Unique Last Name Address Phone Guardian First Name Medicaid Number Birth File Number SSN Guardian SSN Middle Name	

RULE #		FIELDS	
		Column 1	Column 2
160	Birth Date and First Name and at least one of the following fields from column 1.	Address Phone Guardian SSN	
161	Birth Date and Nickname and Gender and at least one of the following fields from column 1.	Address Phone Guardian SSN	
190	Birth Date and First Name and at least two fields from column 1.	Middle Name Guardian First Name Unique Last Name	
191	Birth Date and Nickname and Gender and at least two fields from column 1.	Middle Name Guardian First Name Unique Last Name	
200	First Name, Middle Initial, Last Name, Approximate Birth Date and at least one of the fields from column 1.	Address Phone Guardian SSN	
201	First Name, Middle Initial, Last Name, Approximate Birth Date and at least one field from column 1.	Guardian First Name Unique Last Name	
202	First Name, Last Name, Approximate Birth Date and at least one field from column1 and column 2.	Address Phone Guardian SSN	Guardian First Name Unique Last Name

RULE #		FIELDS	
		Column 1	Column 2
203	First Name, Middle Name, Approximate Birth Date and at least one field from column 1 and column 2.	Address Phone Guardian SSN	Guardian First Name Unique Last Name
204	First Name, Last Name, and approximate Birth Date and at least two of the following from column 1	Address Phone Guardian SSN	
205	First Name, Last Name, and approximate Birth Date and unique Last Name has at least two matches		

Additionally, there are rules to determine if the patients in the subset are "Possible Matches." The "Possible Match" rules are not examined until all the "Merge" rules for definite matches have been examined. Any patients remaining are then examined to see if they are a "Possible Match."

RULES FOR POSSIBLE MATCHES

The "**Possible Match**" rules are not examined until all the "**Match**" (Merge) rules for definite matches have been examined.

Notes: The state can optionally define a list of IRMS's whose patients should never be sent to manual deduplication. If the patient is a "possible" match with any patient that is currently in the registry, the incoming patient record will be deleted rather than sent to manual deduplication. The list of IRMS's must be provided and sent to STC's Help Desk.

Table 0-4: Possible Match Rules

RULE #	DESCRIPTION	FIELDS	
		Column 1	Column 2
These rules do not have "rule" numbers.			
NONE	First Name, Middle Initial, Last Name, Birth Date.		
NONE	First Name, Middle Name, Last Initial, Birth Date.		
NONE	SSN		
NONE	Medicaid Number		
NONE	Birth File Number		
NONE	At least one field from column 1 and column 2.	First Name Birth Date	Address Phone Guardian SSN
NONE	First Names are "Like" sounding and Approximate Birth Date matches and at least one field from column 1.	Address Phone Guardian SSN	
NONE	Birth Date and at least two fields from column 1.	First Name Middle Name Guardian First Name Unique Last Name	
NONE	Birth Date and First Names are "Like" sounding and Last Names are "Like" sounding.		

Finally, it is possible that the rules for "Possible Matches" have flagged too many records, so we apply additional "Separate Rules" for special cases. These all apply only to the records that have been flagged as "Possible Matches."

SEPARATE RULES / SPECIAL CASES

It is possible that the rules for "Possible Matches" have flagged too many records. There are additional "separate" rules for special cases.

Notes: These rules do not have "rule" numbers.

Table 0-5: Separate/Special Cases Rules

RULE #	DESCRIPTION	FIELDS	
		Column 1	Column 2
NONE	Birth Dates are very different: Day, Month, Year do not match and Birth Dates are more than 10 years apart.		
NONE	Twins (Birth Dates are the same and Family/Address information is the same If First Names do not sound "alike" and First/Middle/Last Names were not swapped, then the records are "twins" if any of the following fields in column 1 are true.	Genders are opposite SSN are not null and not equal Provider's ID Number is sequential First Initials do not match and the First Names, excluding First Initials do not sound "alike." If both the First and Middle Names are populated and the first three characters of both Names are different.	
NONE	If Birth Date does not match and it is not an Approximate Birth Date and the First Name does not		

RULE #	DESCRIPTION	FIELDS	
		Column 1	Column 2
	match, the First Name does not match as a Nickname.		
NONE	If the Middle Initials are present and different in both names and there are no swapped initials.		
NONE	If Birth Date does not match and it is not an Approximate Birth Date and the First Initial is different and the Last Name does not match.		
NONE	If the Birth Date and the Phone Number match but the SSN, Address, and Last Initial do not and neither the First name or the Nickname match and the First Names do not sound similar (this handles the special case of a Phone Number being reissued).		
None	To handle twins with the same First Name, if the records would have otherwise matched and the Middle Initial is different or the Middle Name is populated and sounds different, then the record is sent to manual review.		

RULE #	DESCRIPTION	FIELDS	
		Column 1	Column 2
None	First Name, Middle Name, and Birth Date match but none of the following match: Last Name, SSN, Mother Maiden Name, Guardian First Name, and Address; then the records do not a match.		
None	Phone Number and First Name match but Birth Date/approximate Birth Date, SSN, Address, Last Initial, Guardian First Name and Middle Initial do not, then the record is not a match (it is otherwise would have been sent to manual). This addresses a corner-case of a Phone Number being reissued to someone with the same First Name.		
None	If Address and First Name match, but Birth Date/approximate Birth Date, SSN, address, Last Initial, Guardian First Name and Middle Initial do not, then the record is not a match (it if otherwise would have been sent to manual). This addresses a corner-case of person with the same First Name moving to an address that someone else had previously.		
None	If a record is a possible match and		

RULE #	DESCRIPTION	FIELDS	
		Column 1	Column 2
	3+ shots match, then the record is automatically processes as an exact match.		

PROBABILISTIC RULES/LOGIC

The Probabilistic Logic offers a scoring technique to determine a deduplication result of a match, non-match, or unsure (do not know). This logic is also referred to "record linkage process."

The logic includes field comparison functions to return the basic matching weights for each record pair that get stored in weight vectors. The weight vectors are then given to a classifier to calculate a matching decision (match, non-match, or possible match).

Agreement and disagreement weights are computed using the M- and U-probabilities.

$$agreement\_weight = \log_2\left(\frac{m\_probability}{u\_probability}\right)$$

$$disagreement\_weight = \log_2\left(\frac{1-m\_probability}{1-u\_probability}\right)$$

The "Exact String Comparator" function compares the two fields given to it and returns the agreement weight if they are the same and the disagreement weight if they differ.

The "Approximate String Comparators" function allows for partial agreement if the strings are not exactly alike, but almost the same, which can be due to typographic and other errors.

All implemented string comparison functions return a value between zero (two strings are completely different) and one (two strings are the same). The end result is a "minimal approximate string similarity measure tolerance" which will be a number between zero and one.

If the two strings are the same (i.e., the similarity measure returned by the approximate string comparator is one), the agreement weight is returned. If the value is less than one, but larger or equal to the "minimal approximate value", then a partial agreement weight is calculated using this formula.

$$partial\_agreement = agree\_wt - \left(\frac{1 - sim\_measure}{1 - \min\_approx\_value} \right) * (agree\_wt + abs(disagree\_wt))$$

After records have been compared and weight vectors calculated, a classification of record pairs is performed—links, non-links, or if the decision should be done via human review—possible links. This classifier simply sums all the weights in a weight map, and then uses two thresholds (lower and upper) to classify a record pair into one of three classes: links, non-links, or possible links. These thresholds are the boundaries that define a definite match or non-match. Anything in between these boundaries will be sent to the System Administrator for human review.

MANUAL DUPLICATE IDENTIFICATION PROCEDURE

This is a "manual" process when an administrator visually reviews both the incoming patient record and the record residing in the "Reserve" table and makes a decision. This decision is either to create a **new** patient record or **merge** the patient records.

More information regarding "Manual Deduplication" can be located in the chapter titled, "Using Manual Deduplication."

AUTOMATIC MASTER DUPLICATION PROCEDURE (SCANNING)

This is an automatic process that reviews all of the records in the Patient "Master" Table. For example:

A patient named "John Aaron Smith" born on 1/1/1999 enters the registry and is displayed in manual review against a patient named "John Aaron Smitty" born on 10/10/1999.

Since these records are not "similar" enough to be safely merged, the person performing the "manual process" decides the records are different; thus, resulting in two separate patient records.

Weeks later, the sending of the "Smitty" record corrects the record and changes it to "John Aaron Smith" born on 1/1/1999. Now the records are similar enough to be merged. Since both records exist in the "Master" tables as two different patients, the "Master De-Duplication" process is used to merge them.

Appendix B.

Alaska has small population, and the incidence of two unique persons having the same first name, last name, and date of birth is very small – less than 0.1% according to our analysis. Therefore VacTrAK Support will merge two records if the first name, last name and date of birth are the same based on the following analysis.

Deduplication Analysis on Permanent Fund Data

A random sample of 50,000 records from the database of Permanent Fund Dividend applications was selected. The records were sorted and duplicates counted. The following numbers were found.

Criteria	Unique	Same	% of Total	% of Duplicates
FLMD	49984	(16)	= 0.032%	- 0.064%
FLD	49984	(16)	= 0.032%	- 0.064%
LD	49752	(248)	= 0.496%	- 0.992%
FLM	49783	(217)	= 0.434%	- 0.868%
FL	48401	(1599)	= 3.198%	- 6.396%

(F = First name; L = Last name; M = Middle name; D = Date of birth)

This means that out of 50,000 records, only 16 shared the first name, last name, and date of birth. This means that if VacTrAK (manual and automatic) merged all patients with the same first and last name, regardless of date of birth, the system would be correct for 92.6% of the merges.

If it merged all patients with the same first and last name and date of birth, the system would be correct for 99.94% of the merges.

> *Note: Upon further review of the 16 duplicates from the PFD, it was apparent that in each case, the records were in fact duplicates, as determined by Social Security number, address, and guardian information. This can be explained by the fact that the PFD data is a collection of applications, and thus when a single person applies twice, then they would appear in the PFD dataset twice. Thus, the duplicate count for the same first and last name and date of birth in the 50,000 record sample is effectively 0.*

Deduplication Analysis a Public Health Nursing Dataset

The same process was done for a Public Health Nursing dataset with 18,574 patients. The results are as follows:

Criteria	Unique	Same	% of Total	% of Duplicates
FLMD	18569	(5)	= 0.027%	- 0.054%
FLD	18567	(7)	= 0.038%	- 0.076%
LD	18430	(144)	= 0.775%	- 1.55 %
FLM	18405	(169)	= 0.910%	- 1.82 %
FL	18128	(456)	= 2.455%	- 4.91 %

(F = First name; L = Last name; M = Middle name; D = Date of birth)

This has a lower incidence of duplication than the 50,000 randomly selected PFD records.